View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PUblication MAnagement

MANUELA SASSI
ILC-CNR
Pisa, Italia
manuela.sassi@ilc.cnr.it

La obra de Alejo Carpentier en versión digital: historial, descripción y propuestas[1]

1. Premisa histórica

El Instituto de Lingüística Computacional del Consejo Nacional de Investigaciones (CNR) fue creado en 1978 como órgano independiente a partir del Centro de Cálculo Eléctronico de Pisa (CNUCE), del que formaba parte desde los años 60.

Justo a partir de esa época, un grupo de investigadores y programadores comenzó a trabajar con instrumentos informáticos sobre temas de lingüística, continuando los enseñamientos de Padre Busa, el iniciador de esta disciplina en Italia [1].

Una de las primeras investigaciones fue el estudio de la Divina Comedia de Dante Alighieri [2]; a partir de ésta, se continuó - en colaboración con la Academia de la lengua italiana (denominada Crusca [3]) - con el repertorio de todas las obras escritas en italiano desde sus orígenes. De esta manera, después de 25 años, fueron totalmente digitalizadas, revisadas y en camino para la lematización a través de un procedimiento con el cual se ha ido incrementando un diccionario de formas antiguas del italiano, enriqueciéndose con la incorporación de obras lematizadas. La lematización se realizó con criterios selectivos, o sea, se seleccionaron los lemas sobre la base de su importancia para la historia del idioma y se eligieron los contextos más significativos.

A partir de los años 70, se realizó el *Dizionario di Macchina dell'Italiano* (DMI) [4], o sea la sistematización del diccionario del idioma italiano, compuesto por unos 80.000 lemas y 130.000 definiciones estructuradas, codificadas y sintetizadas; además se estudiaron e implementaron algoritmos informáticos con los cuales se pudieron elaborar todas las formas a partir de los lemas, de sus categorías y de las reglas de conjugación, calculadas en alrededor de 1.500.000 formas.

Estos y otros resultados permitieron que el ILC se ganara su independencia y que se estableciera la disciplina con el nombre de Lingüística Computacional.

A lo largo de los años 80 se consolidó el procedimiento de tratamiento automatizado de los textos, con el desarrollo de programas que a través de una codificación estándar de base, permitieron que se pudieran respetar las características de los textos en cualquier idioma y en cualquier alfabeto (por ejemplo: se hicieron trabajos con textos griegos, cuneiformes y con inscripciones en árabe antiguo).

Todavía la tecnología no permitía una representación adecuada de los fenómenos lingüísticos: lentamente pasamos de las palabras en mayúscula (en pantallas e impresoras), a la capacidad de reproducir de manera exacta el texto, incluyendo los signos diacríticos. En aquellos tiempos hubo una evolución muy lenta, comparada con los avances de la tecnología de hoy; además todo esto funcionaba en los mainframes, esas inmensas computadoras que tomaban el espacio de una catedral y cuya gestión estaba en manos de pocas personas que eran los verdaderos dueños del lenguaje informático.

El resultado de una investigación tardaba años, el usuario tenía que esperar los resultados y si estos no eran los esperados, había que volver a plantear el problema. Esto evidenciaba cómo la interacción entre los informáticos y los filólogos, humanistas, lexicólogos era muy dificil y presentaba múltiples incomprensiones.

Todo esto cambió de manera radical con el avance tecnológico y con la llegada del ordenador personal, que dio un nuevo empuje a estos tipos de aplicaciones de la informática. A mitad de los años 80 el ILC se había dotado de ordenadores personales y se comenzó a convertir todo el procedimiento. Los estudios sobre la codificación se mantuvieron y se armonizaron con los que se estaban desarrollando en todo el mundo.

En esa época ya funcionaba la red académica, nacida de un ramo de la red militar (ARPA [5]), y se usaba el correo electrónico desde las primeras décadas de los años 80, con lo cual se pudieron adelantar muchas investigaciones en colaboración con otras instituciones. Ya el ILC formaba parte de la comunidad científica internacional, en la que se cuidaba mucho la codificación y los estándares (TEI) [6] en el tratamiento de datos textuales, gramaticales, lexicales y de descripción del lenguaje.

Toda la experiencia anterior en el uso de grandes computadoras (mainframe) sirvió como base para el desarrollo del nuevo software denominado Data Base Textual (DBT) [7], así como el patrimonio de textos acumulados en años de trabajo: todo fue trasladado a los nuevos medios informáticos en los primeros años de 1990. Este trabajo terminó en 1995, al abandonar el mainframe, substituido por los ordenadores personales. Este hecho determinó el cambio más importante, porque de esta manera se hizo posible que cada estudioso pudiera desarrollar su propio camino de investigación con autonomía.

---

[1] Se han revisado los textos, mas por la calidad de la impresión de algunos de ellos, ha sido bastante largo el trabajo y seguramente no lo suficiente. Por esto mis agradecimientos van a Paolo Picchi, por su duro trabajo y sin él no hubiera sido posible agregar los textos de este corpus. Se agradece Lilia Carpentier, por su confianza en el trabajo y su ayuda. Espero de verdad que el trabajo pueda ser continuado con los demás textos, que, si alguna Editorial tenga ya en forma digital, pudieran agregarse al corpus, con mucho menos trabajo.

2. DBT y sus funciones

Esta experiencia previa permitió que el DBT mantuviera todas las funcionalidades de los procedimientos anteriores y que el pasaje a las nuevas tecnologías fuera más facil, sin pérdida de tiempo ni de datos, transformándose en archivos del DBT.

Además se enriqueció de nuevas funciones antes impensables como, por ejemplo, la posibilidad de consultar conjuntamente archivos inmensos - se trata de millones y millones de palabras - a través de un diccionario con thesauro incluido (esto para el italiano, pero se está terminando para el español, el francés, el latín y el árabe).

La producción literaria más importante de la Historia de la lengua italiana se encuentra en un CD-rom, que se comercializa en Italia y que se puede consultar a través del DBT [8].

En la última versión se puede interrogar por lema (en italiano), evidenciando todas las formas presentes en el corpus, con sus contextos, pudiéndose consultar además, por variantes (1 o 2 carácteres). Se pueden calcular automáticamente los grupos de palabras repetidos; se puede extraer la incidencia, a partir de una o más palabras, con aquellas que se encuentran en sus contextos, eligiendo una distancia máxima o el entorno deseado (antes, después o ambos).

En el caso de textos antiguos, que son depositarios de la historia de la lengua - a partir de la lematización, que puede ser asistida o semi-automática - se puede consultar el texto por categoría gramatical, por palabra o por lema.

Todo esto permite al estudioso de encontrar los datos necesarios para sus investigaciones de manera muy sencilla. A través de botones que sugieren la función correspondiente, se extraen las informaciones básicas, o sea: si ciertas palabras aparecen en el texto, cuántas hay, dónde están y qué entorno tienen; a partir de ahí se pueden buscar más palabras juntas, o averiguar si existen palabras distintas que pertenecen a la misma raíz.

Los contextos de las palabras se visualizan rápidamente, independientemente del tamaño del corpus que se está analizando. El orden normal de presentación es el que aparece en el texto, pero se puede seleccionar otro orden, por ejemplo: en el caso de un verbo, eligiendo el orden alfábetico de las palabras sucesivas (orden derecho) se puede indagar sobre la regencia de éste. Esta función resulta muy útil porque todos los contextos se visualizan ordenados y la palabra clave aparece en el centro de la frase destacando muy bien las construcciones que aparecen en el texto.

Otra posibilidad de la función de extraer contextos es que el usuario puede modificar su extensión en dependencia del tipo de análisis que está haciendo; este parámetro se puede variar en cualquier momento y queda activo hasta su sucesiva variación. Para expresar la extensión del contexto se usa un valor numérico que se corresponde con la cantidad de elementos del texto antes y después de la palabra clave, incluyendo en esta cuenta los eventuales signos de puntuación.

Existe también una versión que permite la consulta de archivos textuales a través de Internet [9]. Se puede utilizar el mismo archivo, elaborado por el DBT estandar, con las funciones principales, obviamente sin poder modificar los archivos, que asì se encuentran al alcance de quien desee consultarlos [10].

3. El DBT y el corpus de Alejo Carpentier

Hace algunos años se comenzò la digitalización de la obra de Carpentier y todo eso tomó camino en 1998 a partir de un encuentro con Lidia en la Fundación Alejo Carpentier, durante el cual me donó un ejemplar de la edición princeps de *El reino de este mundo*. La digitalización se llevó a cabo sin ayuda económica, mi hijo Paolo se hizo cargo del proceso de escanerización. El proyecto del corpus ha sido presentato en el 2006 al *XXVIII Convegno Internazionale di Americanistica* en Perugia (Italia) [11].

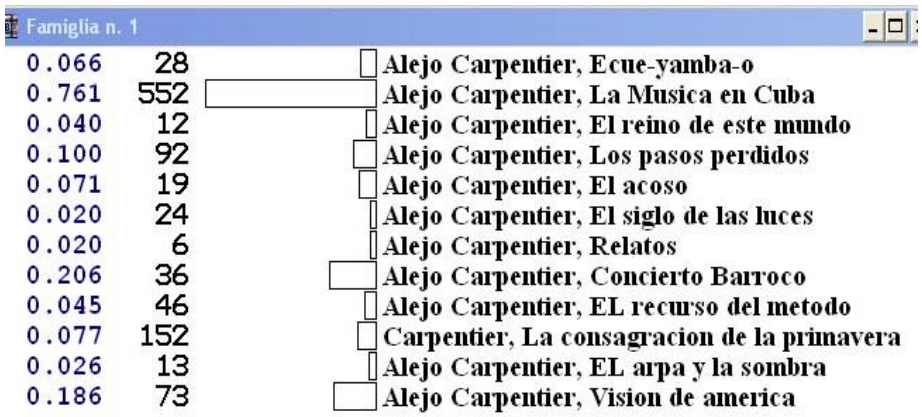
El corpus se compone de los textos que se elencan en la tabla siguiente, donde se encuentran las informaciones sobre cada uno de los libros que se utilizaron para la digitalizacion con el escaner, incluso la sigla que los identifica en el corpus. Hay que señalar que los 5 relatos breves se encuentran todos en el mismo libro, por lo tanto se habla de 14 libros, pero se cuentan 18 textos.

Sigla	Textos	Año	Edición	N.Palabras
ACO	El acoso	1956		26.597
ADV	Los advertidos	(1974?)	UNEAC, 1974	4.398
ARPA	El arpa y la sombra	1978	Ediciones UNION, 1992	49.759
CAM	El camino de Santiago	1967	UNEAC, 1974	12.252
CB	Concierto barroco	1974	Editorial Letras Cubanas, 1979	17.518
CON	La consagración de la primavera	1978	Editorial Letras Cubanas, 2001	197.989
DAS	El derecho de asilo	1972	Editorial Arte y Literatura, 1976	8.288
EYO	¡Ecué-yamba-ó!	1933	Ed. Arte y Literatura, 1977	42.244
FUG	Los fugitivos	(1974?)	UNEAC, 1974	3.721
LUZ	El siglo de las luces	1962		118.398
MEM	Páginas de memoria	2004		73.160
MUS	La música en Cuba	1946	Ed. Pueblo y Educacion (1989)	2.460
PAS	Los pasos perdidos	1953	Alianza Editorial, 1998	91.693
REC	El recurso del método	1974	Siglo XXI Editores, 1974	102.697
REI	El reino de este mundo	1949	Ed. princeps (facsimilar) 1999	29.788
SEM	Semejante a la noche	1958	UNEAC, 1974	4.418
VDA	Visión de América			39.415
VIAJE	Viaje a la semilla	1944	UNEAC, 1974	4.645

Los 14 libros escaneados hasta la fecha, son el resultado de una colaboración familiar, donde nos acomuna la pasión por la obra de Alejo Carpentier: primero somos lectores y después exploradores de su obra.

Para dar fe de cuanto significase la *música* para nuestro Autor, les mostramos las frecuencias de ésta y de sus palabras derivadas, que en total son 1052 ocurrencias; el número inicial muestra la frecuencia en el texto de *La musica en Cuba* y entre parentésis se representa la cantidad de textos que contienen las palabras:

1) extramusical	-- (1)
2) music	2 (3)
3)*música	248 (12)
4) musical	67 (7)
5) musicales	33 (7)
6) musical-estético	-- (1)
7) musicalidad	1 (2)
8) musicalizada	-- (1)
9) musicalmente	3 (1)
10) músicas	8 (11)
11) músico	120 (8)
12) musicología	1 (3)
13) musicológicas	-- (1)
14) musicólogos	1 (1)
15) músicos	68 (11)



4. Aplicaciones prácticas

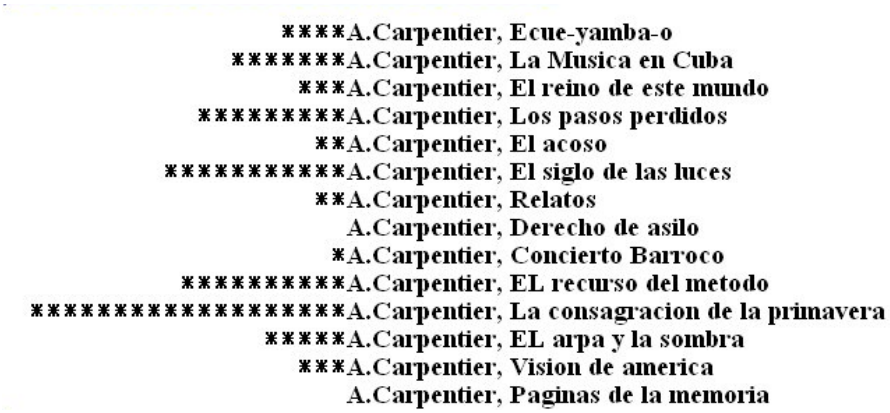
Una de las funciones más útiles para el análisis del texto, que permite hacer evaluaciones inmediatas sobre un determinado campo semántico, es la *coocurrencia*. Sencillamente se trata de la aplicación de una fórmula estadística que calcula la importancia del entorno de una o más palabras claves. Para explicar mejor les mostramos un ejemplo extraído de un archivo en DBT que contiene textos de José Martí, en particular, algunos artículos de periodismo de crítica de arte sobre Europa. En su conjunto son 54.915 ocurrencias con 11.300 formas.

Se indicó *amor** en la ventanilla de solicitud; como resultado se obtuvieron las siguientes palabras (con las correspondientes frecuencias entre paréntesis): *amor (51), amores (20), amoríos (3), amorosa (1), amorosas (1), amoroso (1),* para un total de 77 ocurrencias; a éstas se les aplicó la función de *coocurrencia* seleccionando los siguientes parámetros: 4 palabras antes, palabras 4 después y 2 de frecuencia mínima. El resultado destacó las siguientes palabras en orden de importancia: *respetuoso, mansas, tierno, osado, delicado, tranquilo, primero, vivo, noble, esposa, alma, madre, todo, forma,* etc.

En el cálculo se aplica una fórmula estadística, conocida como *Mutual Information*, la cual se basa en los siguientes elementos: **frecuencia** en el entorno de la palabra dada, que se relaciona con la frecuencia total en el texto, y la **distancia** de la palabra dada.

Esta función fue aplicada al corpus compuesto por los 18 textos de Alejo Carpentier, para un total de **829.454** ocurrencias. Primero se realizó la misma búsqueda de *amor**, de la cual resultó la siguiente lista de palabras, que en total son 146, representada en el gráfico sucesivo:

1) amor	41 (12)
2)*amor-combinatorio	1 (1)
3) amor-de-amantes	1 (1)
4) amor-diversión	1 (1)
5) amores	3 (8)
6) amorío	-- (1)
7) amoríos	-- (1)
8) amor-juego	1 (1)
9) amor-mito	1 (1)
10) amor-muerte	1 (1)
11) amorosa	3 (3)
12) amorosas	1 (2)
13) amoroso	1 (5)
14) amorosos	-- (1)



En este gráfico se da cuenta de las palabras de cada libro y se calcula la relación entre esta cantidad y la frecuencia absoluta de las palabras recuperadas con la búsqueda anterior. A esta lista se aplica la función de coocurrencias estadísticas del DBT, con los mismos parámetros del ejemplo de Martì y el resultado se muestra en la figura siguiente.

Co-Occorrenze statistiche

1)	14	111	9.482	2.643	amor
2)	2	1	13.469	3.000	amor-diversión
3)	2	2	12.469	1.500	arrebatado
4)	2	25	8.825	1.000	brujo
5)	2	12	9.884	3.000	conocimos
6)	2	165	6.103	2.000	danza
7)	2	345	5.039	3.500	dice
8)	2	22	9.010	3.000	duque
9)	2	3	11.884	3.000	eternas
10)	3	22	9.595	1.000	físico
11)	2	942	3.590	3.000	gran
12)	4	430	5.721	2.000	hacer
13)	3	11	10.595	2.000	hicimos
14)	2	258	5.458	3.000	muerte
15)	2	120	6.562	2.000	necesario
16)	2	502	4.498	2.000	negro
17)	2	252	5.492	3.000	pesar
18)	2	41	8.112	1.500	plano
19)	4	26	9.769	1.250	sublime
20)	2	532	4.414	2.500	tenía
21)	2	273	5.376	4.000	tiempos
22)	2	5	11.147	4.000	títere
23)	2	1059	3.421	3.000	vez
24)	3	123	7.112	2.667	ópera

Con estos dos ejemplos podemos apreciar la diversidad de los temas que se relacionan con la misma palabra en los dos autores.
Con esta ejemplificación se desea dar una idea de las posibilidades que ofrece esta herramienta, sin pretensión alguna de análisis profundo de los textos citados, que estaría a cargo de los estudiosos que pueden aprovechar del instrumento según sus necesidades.
Todas las funciones son interactivas, o sea, se hace una búsqueda y el sistema visualiza de inmediato la respuesta. Existen otras modalidaes que se pueden aplicar a un archivo determinado, donde los resultados se registran en un fichero externo indicado por el usuario, pudiéndose utilizar después con un editor de texto o con el mismo Word.
Se pueden realizar además las concordancias de un texto, de un corpus o de una lista de palabras escogidas, que el DBT organizará en un formato tipográfico.
Los índices que se pueden obtener son: frecuencia en orden alfabético, frecuencia descreciente (véase *anexo 1*), índice inverso (a partir de las finales de palabras), index locorum, incipit, excipit, secuencias de caracteres y secuencias de palabras; esta última es una función automatizada que denota la existencia de las secuencias repetidas de palabras a lo largo de un texto determinado. Por ejemplo: aplicamos esta funcion (secuencias de palabras) a los primeros 4 textos digitalizados, del resultado que se obtuvo se eligieron algunos sintagmas que aparecen en el *anexo 2* (después de haber descartado los compuestos de puras palabras funcionales, personajes o sitios de la obra). Se subraya que practicamente no hay frases que se repitan en los 4 textos, escluyendo *de este mundo* (en 4) y *reino de este mundo* (en 3); lo que demuestra que Alejo fue un gran inventor de imágenes nuevas. Con su increíble capacidad creadora y con su gran riqueza de vocabulario nunca tuvo la necesidad de repetirse. Sin embargo se destacan estas locuciones adverbiales por ser las más frecuentes:

a la luz	*a pesar de*	*en busca de*	*en todas partes*
a la sombra	*a través de*	*en cuanto a*	*en torno a*
a la vez	*al cabo de*	*en la noche*	*junto a*
a lo largo	*al pié de*	*en medio de*	*por vez primera*
a modo de	*cada vez más*	*en tanto que*	*una suerte de*

5.Concordancias

Como antes dicho, con el DBT se pueden obtener de manera automática las concordancias de las formas, que el sistema produce como fichero, listo para imprimir con sus características tipográficas. El siguiente es un ejemplo de concordancia de la palabra *Haití* en el corpus de Alejo Carpentier:

Haití *38*

1	que los braceros de	**Haití** aceptaban jornales	*- EYO-14Pag.014*
2	había sido llevada a	**Haití** por su padre.	*- EYO-21Pag.076*
3	agradables que los de	**Haití** », para explicar el	*- MUS.Cap.I Pag.026*
4	fueron a buscar a	**Haití** , para forjarse una	*- MUS.Cap.I Pag.026*
5	en Cuba. en	**Haití** , en el Brasil	*- MUS.Cap.I Pag.035*
6	*rada* (	**Haití**) y aparece,	*- MUS.Cap.VI Pag.118*
7	Director del Conservatorio de	**Haití** » (Guillen),	*- MUS.Cap.VII Pag.135*
8	Liberados los esclavos de	**Haití** , abolida la trata	*- MUS.Cap.VII Pag.137*
9	*La duquesa de*	**Haití** y otras zarzuelas	*- MUS.Cap.XIII Pag.229*
10	_aún viviente en	**Haití** . donde una	*- MUS.CapXVIPag.267*
11	la música que en	**Haití** se llama *vodú*	*- MUS.Cap.XVIPag.270*
12	de las tierras de	**Haití** , de haber hallado	*- REI-PrologoPag.0008*
13	durante mi permanencia en	**Haití** , al hallarme en	*- REI-PrologoPag.0012*
14	era privilegio único de	**Haití** , sino patrimonio	*- REI-PrologoPag.0013*
15	Estado, Rey de	**Haití** , Soberano de las	*- REI-3.06Pag.0159*
16	del primer rey de	**Haití** .	*- REI-3.07Pag.0168*
17	sin detenerse en	**Haití** . Víctor, muy	*- LUZ-Cap.1-XIPag.282*
18	de ultramar. En	**Haití** , lo hicieron	*- LUZ-Cap.4-XXXIIPag.282*
19	Ahora los negros de	**Haití** quieren su	*- LUZ-Cap.4-XXXIIPag.282*
20	los tambores tronaban en	**Haití** : En la región	*- LUZ-Cap.4-XXXIIPag.282*
21	que los negros de	**Haití** se negaron a	*- LUZ-Cap.4-XXXIIPag.282*
22	ejemplo de los de	**Haití** . Cargaban los	*- LUZ-Cap.5-XLPag.282*
23	entonces; como en	**Haití** , cazando negros;	*- REC-17Pag.268*
24	" -"Como en	**Haití** " -dije.	*- REC-18Pag.281*

25	acaso años : mire	**Haití**	donde, pasándose del	- *REC-18Pag.282*
26	sobre el Vodú de	**Haití**	que se refería a	- *CON-21Pag.0223*
27	cónsul de Inglaterra en	**Haití**	, antes de trasladarse	- *VDA-3Pag.035*
28	*abakuá*), en	**Haití**	(los vevés trazados	- *VDA-26Pag.135*
29	*vodú* de	**Haití**	, son auténticas	- *VDA-26Pag.137*
30	guerras de independencia de	**Haití**	-con la admirable	- *VDA-26Pag.139*
31	de las Antillas y	**Haití**	una auténtica escuela	- *VDA-26Pag.140*
32	República Dominicana,	**Haití**	, Puerto Rico,	- *VDA-27Pag.143*
33	Paulina Bonaparte, en	**Haití**	, el mariscal	- *VDA-27Pag.146*
34	Santo Domingo, hoy	**Haití**	, y juraron proclamar	- *VDA-27Pag.150*
35	reclamaban los negros de	**Haití**	-precursores en esto	- *VDA-27Pag.151*
36	de las revueltas de	**Haití**	, que fueron seguidas	- *VDA-27Pag.151*
37	el libertador de	**Haití**	. Petión, presidente	- *VDA-27Pag.154*
38	Petión, presidente de	**Haití**	, fue aquel que	- *VDA-27Pag.154*

Otro ejemplo de elaboración textual es la búsqueda de la Onomástica: el DBT produce un listado en orden alfábetico. El programa selecciona todas la palabras con inicial mayúscula y produce un índice-borrador, en el cual se señalan con asteriscos aquellas que estén relacionadas con los signos de puntuación o que aparezcan con y sin mayúscula. A partir de este índice, se pueden examinar los contextos relativos a las palabras que generen dudas a través de la interrogación del Corpus, además de averiguar la razón de la inicial mayúscula: si se trata de nombre propio, si se encuentra al inicio de una frase, o si el Autor está dando énfasis a un sustantivo común.
Las palabras siguientes son las primeras del listado alfabético del corpus hasta ahora analizado: la columna 1 representa la frecuencia, la columna 2 representa la presencia o menos de la inicial mayúscula, la columna 3 representa el numero de veces que la mayúscula esta precedida por un signo de puntuación fuerte. En esta lista se han seleccionado las palabras con la frecuencia más alta y que nunca aparecen en minúscula (col.2 = 0).

306	0	0	Habana	74	0	3	Calixto	43	0	1	Batista
304	0	60	Víctor	73	0	1	Madrid	43	0	0	Orinoco
299	0	65	Sofía	68	0	11	Cervantes	42	0	2	Italia
296	0	10	Cuba	68	0	3	JeanClaude	40	0	0	Lenormand
286	0	2	América	63	0	0	Guadalupe	40	0	10	Longina
260	0	20	Juan	63	0	7	Mirta	40	0	0	Mezy
232	0	14	Antonio	62	0	1	Cristo	40	0	3	Miguel
218	0	67	Menegildo	59	0	7	Ogé	40	0	13	Usebio
185	0	7	José	56	0	4	Martínez	39	0	0	Federico
171	0	1	Santiago	56	0	1	Paris	38	0	0	Haití
169	0	3	España	55	0	4	Jorge	38	0	8	Ruth
169	0	3	Francia	53	0	1	Venezuela	38	0	0	Sevilla
164	0	1	Europa	52	0	0	Córdoba	37	0	1	Inglaterra
156	0	10	Enrique	52	0	6	Ofelia	37	0	0	Londres
143	0	9	Peralta	52	0	2	Rusia	36	0	1	García
141	0	16	Carlos	52	0	3	Saumell	35	0	2	Alemania
137	0	9	Gaspar	51	0	0	Antillas	35	0	0	Brasil
133	0	3	México	51	0	1	Cayena	35	0	0	Curador
122	0	0	Pedro	51	0	0	Cristóbal	35	0	2	Montezuma
121	0	32	Mouche	51	0	7	Espadero	35	0	0	Saint
114	0	9	Teresa	50	0	3	Filomeno	34	0	5	Amaliwak
112	0	1	Hugues	50	0	9	Mackandal	34	0	0	Elmira
95	0	0	Noel	49	0	2	Hoffmann	34	0	5	Henri
93	0	1	María	48	0	4	Christophe	34	0	0	Mendoza
89	0	1	Colón	48	0	1	Francisco	33	0	1	Beethoven
75	0	1	Luis	44	0	6	Roldan	33	0	0	Castro

6. Conclusiones

Con estas páginas se quiere dar una muestra mínima de las innumerables aplicaciones del DBT para la creación y la difusión de los recursos lingüísticos que puedan ser útiles a los estudiosos de Literatura, de Lingüística y de todas las disciplinas con respecto a su lenguaje especializado.
Los resultados que se obtienen con la funciones del DBT pueden ser usados directamente o elaborados en pasajes succesivos para producir páginas web, dando así herramientas a los estudiosos sin tropezar con el problema del derecho de autor, ya que los textos completos estan criptados en la memoria, y se pueden sólo visualizar los contextos requeridos, los índices de palabras o los resultados de las funciones de búsqueda.
A través de Internet es posible poner al alcance de los estudiosos la posibilidad de consultar los archivos y de extraer las concordancias de los términos deseados (siempre y cuando se haya la autorización necesaria para hacerlo) dado que se trata de una extracción de información y no de la duplicación de textos, que permanecen protegidos en la base de datos.
Como ejemplo de uso de base de datos textuales de autor y de recursos computacionales puestos a disposición en Internet, véase el sitio web dedicado a Carlo Emilio Gadda [12], prestigioso autor italiano, al cuidado de la autora e de la investigadora Maria Luigia Ceccotti [13].

7. Anexo 1

En el siguiente listado se registran los sustantivos, adjetivos y verbos que se encuentran en el formario completo del corpus de Alejo Carpentier, extraídas de entre las primeras 500 palabras más frecuentes.
La primera columna es el total de los libros y la segunda la frecuencia total:

14	792	día	13	271	cabeza	13	200	trabajo	12	155	sombras
14	770	noche	14	271	hora	13	199	sangre	13	154	madera
14	740	mundo	13	269	camino	13	198	estilo	12	153	palacio
14	730	casa	13	269	momento	12	197	fondo	12	153	paredes
14	709	hombre	12	268	decir	13	195	tener	11	153	plaza
13	683	hombres	14	266	calle	13	194	silencio	12	153	sentido
13	674	bien	7	264	magistrado	13	191	pensar	13	151	espíritu
14	666	tiempo	13	260	presencia	12	191	voces	12	151	fuerza
14	640	días	14	259	cosa	14	186	mesa	14	151	recuerdo
14	589	ciudad	13	258	muerte	13	184	meses	9	150	continente
14	575	años	13	258	puerta	12	183	pasar	12	150	negra
12	524	música	13	258	saber	14	182	cuenta	12	150	orden
13	506	tierra	14	257	gentes	14	181	iglesia	12	149	fuego
13	502	negro	13	256	cabo	13	181	poder	14	149	niños
14	431	ojos	14	256	viaje	14	180	llegar	14	146	hojas
14	424	vida	12	252	pesar	14	180	ritmo	11	145	baile
13	423	gente	13	251	año	11	179	árboles	11	145	selva
14	415	mano	13	250	hablar	13	179	idea	14	144	caer
13	412	fin	11	247	parece	13	179	señor	11	143	existencia
12	406	negros	14	247	pie	12	178	indios	13	143	patio
12	401	guerra	13	244	padre	14	178	piedra	13	142	francés
13	398	cosas	11	243	teatro	13	177	suerte	10	142	gobierno
13	386	voz	14	242	nombre	13	176	luces	13	142	real
13	379	mañana	13	240	cara	13	175	época	10	142	río
14	363	mujer	13	239	pasado	13	173	santo	7	141	cubano
13	356	agua	14	236	calles	11	172	tono	14	141	noches
14	355	ver	13	236	dios	8	171	santiago	10	141	rosario
14	354	manos	13	230	suelo	11	170	miedo	12	139	ayer
13	347	lugar	12	229	obra	12	170	rey	12	139	cambio
13	340	tarde	14	228	luz	12	169	españa	10	139	perro
12	327	isla	14	228	madre	12	169	francia	13	138	hijo
14	324	mujeres	12	228	palabra	13	169	pueblo	12	138	libro
14	324	paso	13	227	verdad	11	168	olor	11	138	músicos
13	316	mar	13	226	papel	10	167	orquesta	12	137	caminos
13	308	oro	12	223	boca	13	166	blanco	13	137	color
14	306	habana	12	223	realidad	12	166	vieja	11	137	estar
14	306	horas	14	220	casas	12	165	danza	4	137	gaspar
8	304	víctor	11	218	joven	14	165	dar	13	137	puertas
13	302	parís	1	218	menegildo	13	164	aguas	13	137	virgen
14	300	historia	14	217	viejo	12	164	europa	13	135	salir
2	299	sofía	13	216	carne	10	163	presidente	14	135	verde
12	298	cuerpo	13	213	gesto	13	162	negras	12	134	naves
10	296	cuba	12	213	santa	12	162	presente	13	133	andar
13	290	palabras	11	212	revolución	8	161	músico	7	133	cubana
12	286	américa	12	211	cielo	11	161	obras	10	133	méxico
13	286	hoy	13	208	caso	14	161	puerto	12	132	calor
12	286	país	12	206	aire	12	160	libros	13	132	falta
13	281	siglo	13	205	espera	13	160	sombra	12	132	familia
14	280	sol	14	204	estado	13	158	brazos			
14	273	tiempos	12	204	tierras	11	156	acción			

8. Anexo 2

En la siguiente tabla aparecen:
A número de textos en que se encuentra el sintagma
B frecuencia total del sintagma
ACO=El acoso,
LUZ=Siglo de las luces,
PAS=Los pasos perdidos,
REI=El reino de este mundo

con la frecuencia del correspondiente sintagma en la obra.

A	B	ACO	LUZ	PAS	REI	
1	5			5		*a ambos lados*
1	4		4			*a diestro y siniestro*
1	4		4			*abolición de la esclavitud*
1	5	5				*academia de corte*
1	10		10			*agente del directorio*
1	3			3		*bajo la lluvia*
1	4			4		*bastón de ritmo*
1	4			4		*botella de aguardiente*
1	3		3			*botella de tafia*
1	5		5			*botella de vino*
2	18		12	6		*casa de gobierno*
1	3	3				*con los ojos dormidos*
1	3		3			*coronado de espinas*
1	5	5				*cárcel de mujeres*
2	6		3	3		*de esta(s) tierra(s)*
4	10	1	2	4	3	*de este mundo*
3	15	7	5	3		*de la cruz*
3	17		11	3	3	*de otros tiempos*
1	3				3	*explanado de honor*
1	3			3		*gesto de denegación*
1	7				7	*gorro del obispo*
1	3			3		*ha fundado una ciudad*
1	6		6			*imperio del norte*
1	4	4				*imprenta de tarjeta de visita*
1	3		3			*ingenios de azucar*
1	8		8			*investido de poderes*
1	4		4			*la felicidad de un hombre*
1	8				8	*la llanura del norte*
1	3		3			*la masa humana*
1	4			4		*la noche de las edades*
1	8			8		*la novena sinfonía*
1	3			3		*la orilla del mar*
3	10		3	4	3	*la otra orilla*
1	4		4			*la revolución francesa*
1	3	3				*la tarde aquella*
1	6		6			*la tierra firme*
1	3			3		*las capas de los árboles*
1	6		6			*llevar la revolución a*
1	3		3			*llevar la revolución a España*
1	3		3			*lo mejor de su tiempo*
1	4			4		*lo necesario para*
2	6		5		1	*los derechos del hombre*
1	3		3			*los días de la pubertad*
1	3			3		*los rincones oscuros*
1	5	5				*los tiempos del tribunal*
1	4		4			*place de la victoire*
1	7		7			*plaza de la victoria*
3	12	3	3	6		*por el camino*
2	7		4	3		*por la humedad*
3	11	3	4	4		*por los caminos*
1	3			3		*quinta de trompas*
3	7		2	2	3	*reino(s) de es(t)e mundo*
1	6		3			*representante del pueblo*
1	3			3		*resmas de papel*
1	4		4			*retrato del incorruptible*
2	9		6	3		*ropas de luto*
2	7	4		3		*sala de concierto*
1	3			3		*santos y señas*
1	5		5			*se encogió de hombros*
1	3			3		*se volvió hacia mí*
1	3		3			*siglo de las luces*
2	10		6	4		*sin hacer caso*
1	3			3		*sobre nuestras cabezas*
2	6		3	3		*sol a sol*
1	3		3			*tiene el derecho de*

3	9		3	3	3	*todo el mundo*
1	3		3			*todo el ámbito del caribe*
1	5		5			*todos los hombres*
1	4			4		*toque de queda*
1	4	4				*tronco más espeso*
1	4				4	*ultima ratio regnum*
1	3			3		*un buen día*
1	3		3			*un contento físico*
1	3		3			*un exceso de palabras*
1	3		3			*un ir y venir*
1	3		3			*un pueblo libre*
1	3			3		*una mazorca de maíz*
1	3				3	*volver la cabeza*
1	3			3		*volvió hacia mí*

8. Referencias

[1] Padre Busa, Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices ed concordantiae, Stoccarda, Frommann Holzboog, Stoccarda, 1974-1980, 56 volumenes de alrededor de 1000 páginas cadauno (62.550 páginas). Es La indicización completa de todas las ocurrencias de cada palabra usada en las obras de Santo Thomás. Esta es su obra principal, disponible en CD-ROM y DVD, 1990

[2] La Divina Commedia : Testo, concordanze, lessici, rimario, indici. Milano, IBM Italia, 1965.

[3] Link al sitio de la Accademia della Crusca: http://www.accademiadellacrusca.it/index.php.

[4] Zampolli A. et alii, Il dizionario di macchina dell'italiano. In: Linguaggi e formalizzazioni, Atti del Convegno Internazionale di Studi, Catania, 17-19 settembre 1976, Roma, Bulzoni, 1979.

[5] Link a Wikipedia: http://it.wikipedia.org/wiki/ARPANET

[6] Link al sitio de la Text Encoding Iniative: http://www.tei-c.org/index.xml.

[7] Picchi E. 1991. D.B.T.: A Textual Data Base System. In L. Cignoni, C. Peters (eds.), Computational Lexicology and Lexicography. Special Issue dedicated to Bernard Quemada. II, Linguistica Computazionale, 7. 177-205.

[8] P. Stoppelli, E. Picchi, LIZ, Zanichelli, Bologna, 1993, 4a Ed. 2000.

[9] Biagini L., Picchi E. INTERNET and DBT. In: M. Gellerstam, J. Järborg, S. Malmgren, K. Norén, L. Rogstrom, C. Rojder Papmehl (eds.), EURALEX '96 Proceedings. Seventh EURALEX International Congress on Lexicography. Göteborg (Sweden), 1996, pp. 47-53.

[10] Este link permite la consulta de algunos archivos de textos legislativos cubanos, como por ejemplo el Codigo de Familia y otros: http://www.ilc.cnr.it/pisystem/demo/demo_dbt/demo_singolo/index.htm.

[11] Sassi M., Martí y Carpentier, voces de la América mestiza: recursos lexicales . Comunicazione presentata al XXVIII Convegno Internazionale di Americanistica, Perugia 3-7 maggio 2006, pp. 521-526. 2006.

[12] M.L. Ceccotti, M. Sassi, L'Archivio elettronico delle opere di Gadda in DBT 2000. In: EJGS Supplement no. 3, EJGS 4/2004. (http://www.arts.ed.ac.uk/italian/gadda/Pages/journal/supp3atti1/supplem3.php).

[13] Link al sitio de Carlo Emilio Gadda: www.ilc.cnr.it/CEG.